Methods in bioinformatics

R programming language

R: Projects assignments
Xmas special

ASSIGNMENT

Teachers

Federico Zambelli
federico.zambelli@unimi.it

Matteo Chiara
matteo.chiara@unimi.it

Rules of the game #1

- You **must** register for one of the available exam dates using the SIFA service.
- Available dates are:
  - 26th Jan 2022 – 15.00
  - 10th Feb 2022 – 15.00
  - 25th Feb 2022 – 10.30
  - 21th Jun 2022 – 15.00
  - 05th Jul 2022 – 15.00
  - 25th Jul 2022 – 15.00
  - 20th Sep 2022 – 15.00

Rules of the game #2

- SIFA will open the registration more or less 15 days before each exam date.
- The room for the exam will be communicated on the Ariel website and a couple of days before the exam.
- According to current regulations all exams should be taken in person. But in special circumstances you can take the exam from remote
  - See : rules
- In general, keep an eye on the Ariel website and the MS-Teams channel for last minute communications.

Rules of the game #3

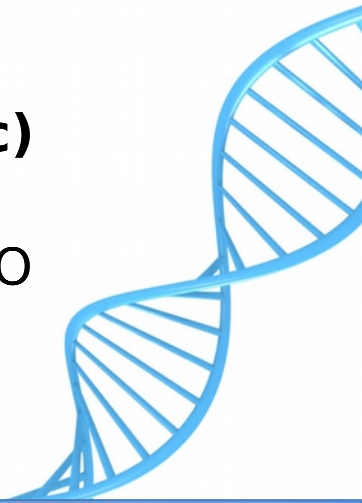
- The first part of the exam consists in producing a (html) report document for one of the available projects
- You can work on a project alone or in group, groups can be composed by two or three students.
  - You are free to choose your partners and assemble groups
- **Reports must be submitted at least 48 hours before the selected exam date**
  - **failing to do so will exclude you from that exam date.**

Rules of the game #4

- Reports must be submitted to both
  - federico.zambelli@unimi.it
  - matteo.chiara@unimi.it
- Reports will be contained in a zip(.zip) archive file.
- The archive **must** contain both the .Rmd and the .html files.
- Additional files that can not be displayed inside the report can be included in the archive,
  - for example image files of Venn Diagrams

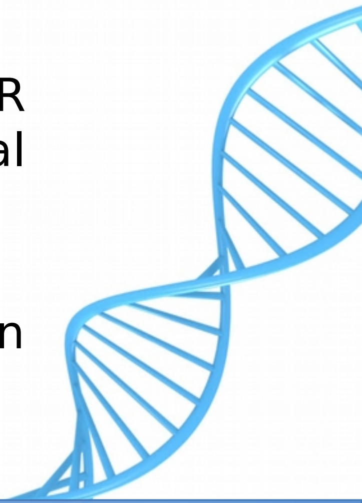
Rules of the game #5

- When you submit a report you must clearly state in your e-mail:
  - **Your name, surname and badge number.**
  - **The name, surname and badge number of ALL the components of your group.**
  - **The project you chose**
    - **The selected date for oral discussion. These can be different for each member of the group (but avoid if possible)**
- All the members of a group **must be put in copy (cc)** when submitting a project report.
- Just to be clear: one submission per group is enough, DO NOT submit the same report for each group member.



Rules of the game #6

- Reports must contain both the code necessary to your analysis and brief explanations of what you are doing, (use comments # for that) why you are doing it, and a **brief discussion on the results**.
- The second part of the exam will consist in an **oral discussion of your report**, including the main findings, and the *interpretation* of the results, followed by questions on concepts we saw during the course.
  - These will include both theoretical aspects of the R programming language and of statistical tests for differential gene expression
  - Discussion will be strictly in English
  - Your answers will help us to assess your individual contribution to the project and general comprehension of the topic .



Rules of the game #7

- You can seek our advice for the project at any moment **before the final submission**, by writing an e-mail and eventually set up an appointment but...
  - No one is going to write the project or any line of code for you.
  - Try to avoid questions that can be easily answered just by looking at the lecture notes and at the many examples you have at your disposal in the walkthroughs.
  - Remember also that you have an help manual for each function and a lot of documentation on the Web.

Rules of the game #8

- If you don’t pass the exam, you will have to resubmit a completely new project, and select an alternative “track”
  - You do not need to split/re-arrange modify the group for the new sumbission, but you may if you want
- If you pass but you are not satisfied with the mark, you can submit a revision of the project where **you MUST address all the critical points** that emerged during the discussion, however:
  - Revisions must be submitted **individually** (not as a group)
  - Revised projects need to be submitted and discussed like any other project
  - Revised projects are not guaranteed to get you a higher grade.
  - **You can revise your project only once**
  - If your revised project is not considered adequate, you will have to submit a completely new project, by selecting an alternative track (see above)

General tips

- You are not studying and practicing R to make me happy but to acquire a powerful tool that could be a key component of your skills set.

- All the projects can be carried out just using what you learned through the course.

- There is no need of concepts / functions / libraries / packages that you do not know (or should know) already.

- This does not mean that you are not free to be curious: if you discover and like some functions or packages that were not covered during the course you can use them,
  - provided that you explain in your report why you did so

Projects: common part

- You will work on a human gene expression profiling RNA-Seq dataset composed by 60 samples from 10 human organs/tissues.
  - Library preparation: **polyA+**
- Data have been preprocessed by us to discard
  - genes with low quality or incosistent annotation
  - Mitochondrial genes
  - tRNAs and rRNAs
- So of the ~ 56k human genes that are annotated in the GenCode annotation only 28.188 high quality genes have been retained

Projects: common part

- For each sample you have the expression values (read counts) for
  - 18805 (high quality) proteing coding genes and
  - 9383 (high quality) non protein coding RNAs (6496 lincRNA, 1771 snRNAs and 1116 miRNAs)
- The dataset consists of 3 files
  - **Counts.csv**: a table containing gene expression values (read counts) for the 28.188 human genes in the 60 replicates
  - **Annot.csv**: a table containing the annotation (gene symbol and class) for the 28.188 genes
  - **Design.csv**: a table containing the experimental design of the RNAseq (i.e the tissue, individual and sex associated with each biological replicate)
  - All the files can be downloaded **here** or from the Ariel website.
  - All files are tab (“\t”) delineated and
    - Have a header line
    - Have row names (genes or samples names) in the first column

Projects: common part

- The dataset is a “cleaned and shrinked” version of the data produced in the context of the GTEX project.
- See *https://gtexportal.org/home/publicationsPage* for a complete list of the publications associated with the GTEX project
  - Try to draw simple but meaningful **biological conclusions** from your analyses and to incorporate them in your report.
  - You are **free to expand** your analyses if you feel engaged to do so.
  - If you are asked to create plots, please give them meaningful titles and labels

Projects:

- The first part of the project is common between all tracks, and consists in the following analysis:
- You need to use the edgeR package in order to
  - 1 read the data into a dgeList object
  - 2 keep only genes that are likely to be expressed (i.e genes that have **more than 10 reads in at least 1 replicate**)
  - 3 perform normalization with *calcNormFactors()*
  - 4 perform a MDS (~PCA) plot of the data
  - 5 select **2 different (and meaningful) biological conditions** and perform a differential expression analysis using the *exactTest()* function
  - 6 create a topTags type of edgeR object containing the list of differentially expressed genes (DEGs)
    - DEGs should have a FDR <= 0.01
  - 7 Assign all the genes into one of the 4 possible classes: DE_UP (FDR<=0.01 and logFC>0), DE_DOWN (FDR<=0.01 and logFC<0), notDE_UP (FDR>0.01 and logFC>0), notDE_DOWN (FDR>0.01 and logFC<0), and then do a **boxplot** of the logFC of the genes belonging to each class

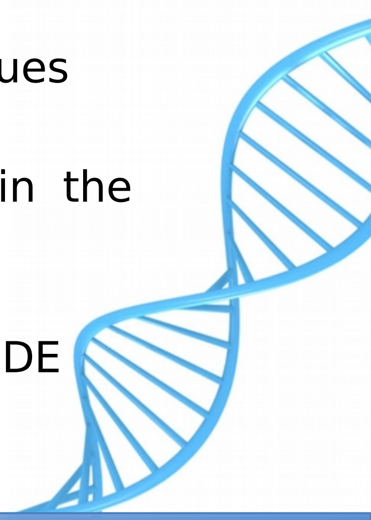
Projects: second part

- For The second part of the project you can select 1 of 3 possible assignments

- **General Tips:**
  - In all the assignments, unless it is explicitly stated not to do so, work only with the genes that are expressed i.e. >= 10 counts in at least 1 replicate
  - Again, **unless** it is explicitly stated otherwise, work always with normalized counts
  - Make always sure that you data tables are **“matched”** (i.e samples should appear in the same order)
  - When plotting use log-scaled values (unless explicitly stated otherwise)
  - If something is not clear, ask clarifications to us
  - “I did not understand the text of the assignment” **will not be considered a valid justification** for failing to do what you are supposed to do

Project #1

- Identify genes showing sex specific expression in the 2 tissues that you considered for the “common part”.
  - Perform a PCA (principal component analysis) to ascertain whether there is separation between biological replicates of different Sexes **(for exery tissue you have 3 individuals of sex 1 and 3 of sex 2)**
  - For each tissue, consider only genes expressed (>10 reads in at least 2 replicates) in that tissue.
  - Use edgeR (exactTest) to perform a differential expression analysis
  - Consider all the genes that show a FDR <= 0.05 as “sex specific” DEGs
- Draw a **Venn Diagram** of the Sex specific genes between the 2 tissues **How many genes are sex specific in both tissues?**
- Finally draw a Venn Diagram between the DEGs (as identified in the common part) and genes showing Sex-specific expression in at least one of the tissues considered.
- How many genes that are DE between the 2 tissues are also DE between the 2 sexes? Do you expect to see many? Why?



Project #2

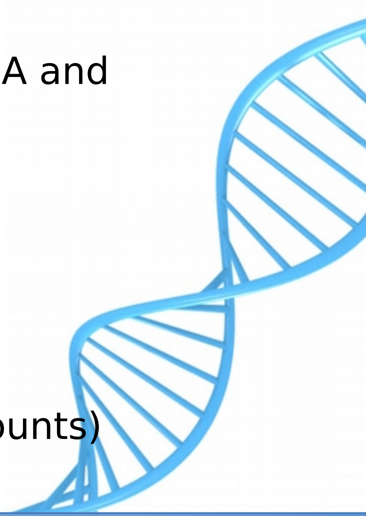
- Identify the **housekeeping genes (HK)**
  - These must have an expression >=10 (read counts) across all the samples.
  - No exceptional change in expression in any single sample:
    - Avg_Exp /2 <= Sample_Exp <= Avg_Exp/2
    - Where Avg_Exp is the mean expression across all the samples and Sample_Exp is the expression in the sample.

- **Consider the DEGs that you obtained in the common part.**
  - Are any of these genes housekeeping according to our definition? How many?
  - Do you expect that many housekeeping genes should be DE? If so why?

- Pick one organ/tissue and draw 5 scatterplots of the **log2(average counts)** of **HK genes** of the tissue you picked, against 5 other tissues of your choice
  - Color genes belonging in different classes (protein coding, lincRNA, snRNA and miRNAs using different colors)
  - Comment the results: are these plots in line with the MDS(PCA)-plot?

- How many of the housekeeping genes are protein coding?
  - How many are lincRNA, snRNA and miRNA?
  - Draw a barplot to illustrate the results of this analysis

- Use boxplots to compare expression values of housekeeping protein coding genes, lincRNAs, snRNAs and miRNAs (use normalized and log scaled counts)
  - Which class of genes is more expressed?



Project #3

- Consider **now another tissue**, different from the two that you have selected in the common part
  - Perform all the (3) possible pairs of differential expression analyses between the 3 tissues that you have selected

- Based on the results of differential expression analyses, classify the genes into one of the following 4 classes: **not DE**, **DE in 1 comparison, DE in 2 comparison** and **DE in all the comparisons.**

- How many genes are DE in one comparison? How many in 2? How many in 3?
  - draw a barplot to illustrate the results of this analysis
  - draw a Venn Diagram to illustrate the results of this analysis

- Create a function that takes in input the ID of a gene and the gene expression counts table.
  - The function must draw a barplot of the mean expression value of the gene in input across the three tissues you selected.
  - Use this function to draw barplots for a 5 of genes in each of the 4 classes of the previous point